

Statistica Sociale

20/03/2025

Per casa (lezione 13/03)

A. Consultare il sito dell'Istat <https://www.istat.it/it/archivio/91926>

farsi una idea delle caratteristiche dell'indagine:

- popolazione di interesse
- indagine campionaria o censimento
- argomenti/temi trattati
- modalità di raccolta dei dati
- ... (altre caratteristiche che vi paiono interessanti)

Per casa (prossima lezione)

A. Consultare il sito dell'Istat <https://www.istat.it/it/archivio/91926>

- farsi una idea delle caratteristiche dell'indagine **MULTISCOPO FAMIGLIE: ASPETTI DELLA VITA QUOTIDIANA**
- popolazione di interesse **famiglie e individui**
- indagine campionaria o censimento **campionaria** (2025: 25.000 famiglie distribuite in circa 800 comuni italiani)
- argomenti/temi trattati **Scuola, lavoro, vita familiare e di relazione, abitazione e zona in cui si vive, tempo libero, partecipazione politica e sociale, salute, stili di vita**
- modalità di raccolta dei dati **CAWI/CAPI** (13/01 - 8/4 2025)
- ... (altre caratteristiche che vi paiono interessanti) banca dati **IstatData**, etc.

MULTISCOPO FAMIGLIE: ASPETTI DELLA VITA QUOTIDIANA (anni vari)

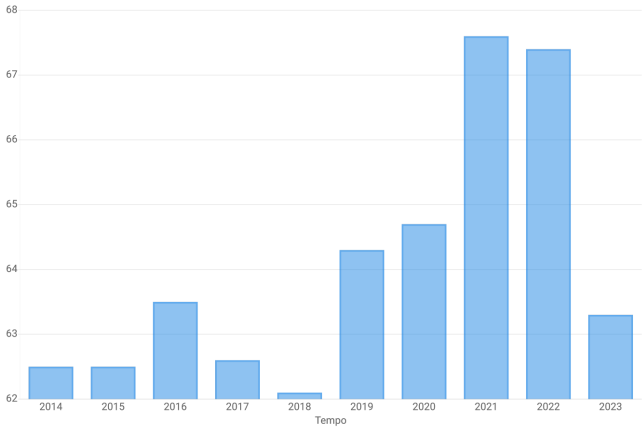
Territorio **Italia** Misura **Per 100 persone con le stesse caratteristiche** Sesso **Totale** Età **18-34 anni**



Tempo	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Indicatore	▲ ▼ ▾	▲ ▼ ▾	▲ ▼ ▾	▲ ▼ ▾	▲ ▼ ▾	▲ ▼ ▾	▲ ▼ ▾	▲ ▼ ▾	▲ ▼ ▾	▲ ▼ ▾
Giovani di 18-34 anni, celibi e nubili, che vivono in famiglia con almeno un genitore	62,5	62,5	63,5	62,6	62,1	64,3	64,7	67,6	67,4	63,3
Giovani di 18-34 anni, celibi e nubili, che vivono in famiglia con almeno un genitore per condizione professionale - occupati	33	31,8	33,2	33,7	38,7	38,2	35,9	37	39,1	39,5
Giovani di 18-34 anni, celibi e nubili, che vivono in famiglia con almeno un genitore per condizione professionale- in cerca di occupazione	28,8	29,7	26,9	24,5	20,9	23,7	24,6	25,2	22,1	20,4
Giovani di 18-34 anni, celibi e nubili, che vivono in famiglia con almeno un genitore per condizione professionale - casalinghe	0,8	0,8	0,7	1,1	1	1,1	1,7	0,8	0,7	0,6
Giovani di 18-34 anni, celibi e nubili, che vivono in famiglia con almeno un genitore per condizione professionale- studenti	35,5	35,5	36,4	37,4	36,1	36,5	36,2	35,6	36,9	38,8
Giovani di 18-34 anni, celibi e nubili, che vivono in famiglia con almeno un genitore per condizione professionale- in altra condizione	1,9	2,1	2,8	3,4	3,2	0,5	1,5	1,3	1,3	0,7

Giovani che vivono in famiglia

Frequenza: Annuale, Territorio: Italia, Indicatore: Giovani di 18-34 anni, celibi e nubili, che vivono in famiglia con almeno un genitore, Misura: Per 100 persone con le stesse caratteristiche, Sesso: Totale, Età: 18-34 anni



Selezione casuale (campione)

I **metodi statistici inferenziali** fanno uso delle statistiche campionarie per fare previsioni sui parametri delle popolazioni

L'**utilità** dell'inferenza dipende da quanto bene il campione rappresenta la popolazione

- È importante ridurre la probabilità di selezionare campioni che per le loro caratteristiche possano **distorcere** la rappresentatività della popolazione portando ad errate conclusioni inferenziali sui valori dei parametri

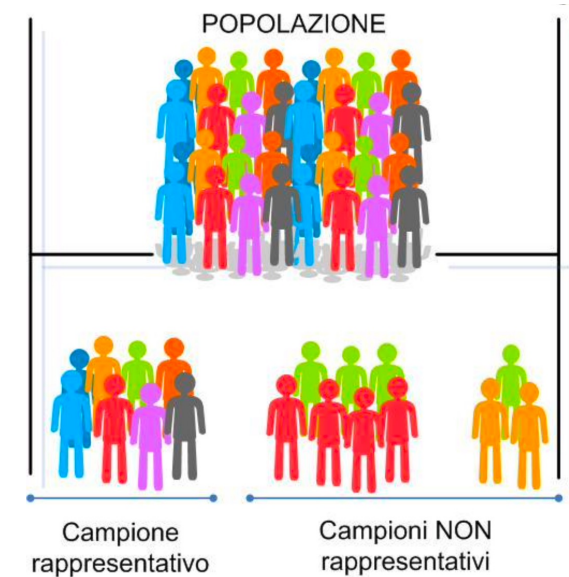
Campione casuale (semplice) di n soggetti estratti da una popolazione è tale se ogni possibile campione di pari numerosità ha uguale probabilità di selezione

- n è la numerosità/dimensione campionaria
- Per selezionare un campione è necessario avere una **lista di campionamento**

Indagine campionaria: selezione di un campione dalla popolazione di riferimento; le informazioni sono raccolte tramite intervista diretta, telefonica o autocompilata (per es., online)

Esperimento: confrontare le risposte sotto diverse condizioni (trattamenti) su cui si ha controllo sperimentale; il piano sperimentale è il processo attraverso cui il ricercatore assegna i soggetti ai diversi trattamenti (in modo casuale)

Studio osservazionale: si osservano i valori delle variabili di interesse ma non si ha controllo sperimentale; non è possibile determinare i rapporti causa-effetto



L'importanza della numerosità campionaria (e della selezione casuale!)

Nel 1936, il settimanale Literary Digest inviò per posta oltre 10 milioni di questionari chiedendo di prevedere chi sarebbe stato il prossimo presidente americano tra Landon e Roosevelt.

Solo il 25% degli intervistati rispose. Il risultato atteso dagli intervistati fu la schiacciante vittoria di Landon.

- Che tipo di campionamento è quello del Literary Digest?
[campionamento volontario]
- Gli intervistati sono stati selezionati da liste di iscritti a riviste, liste di numeri telefonici e liste di appartenenza a club. Sono liste rappresentative della popolazione di riferimento?
[no, rappresentano bene la popolazione della classe sociale medio-alta negli USA del 1936]
- Quale altro problema evidenziate nell'indagine del Literary Digest?
[l'elevato numero di risposte mancanti]

George Gallup ha condotto un'indagine campionaria con lo stesso fine su circa 50.000 persone

Il risultato previsto dagli intervistati da Gallup fu la vittoria di Roosevelt.

Altri metodi di campionamento probabilistici

Campione sistematico: dato un passo di estrazione $k = N/n$, vengono selezionati tutti i soggetti nella lista presenti ogni k soggetti

- Più semplice del campionamento casuale
- Anche se non ha tutte le caratteristiche di un campione casuale, possono essere applicati gli stessi metodi previsti per il campionamento casuale

Campione stratificato: la popolazione viene divisa in strati e da ognuno di questi viene estratto un campione casuale semplice

- **Proporzionale** (se le proporzioni di ogni strato del campione è uguale alla proporzione di popolazione corrispondente al gruppo) o **non-proporzionale**
- Alcune variabili sono più adatte alla definizione degli strati

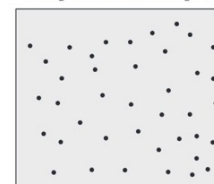
Campione a grappoli: la popolazione viene divisa in molti gruppi (grappoli) ed il campionamento casuale è applicato ai grappoli selezionando tutti i soggetti in essi inclusi

- La maggior parte dei grappoli non viene rappresentata dal campione

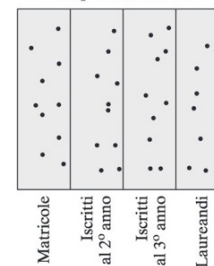
Campione a più stadi: è ottenuto attraverso combinazioni dei metodi precedenti. Ad esempio viene applicato prima un campionamento a grappoli e successivamente vengono campionate delle unità in ogni grappolo selezionato

(per es.: campione di scuole e campione di studenti entro le scuole selezionate)

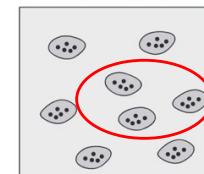
Campione casuale semplice



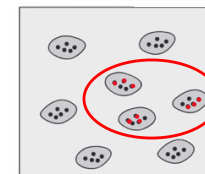
Campione stratificato



Campione a grappoli



Campione a 2 stadi



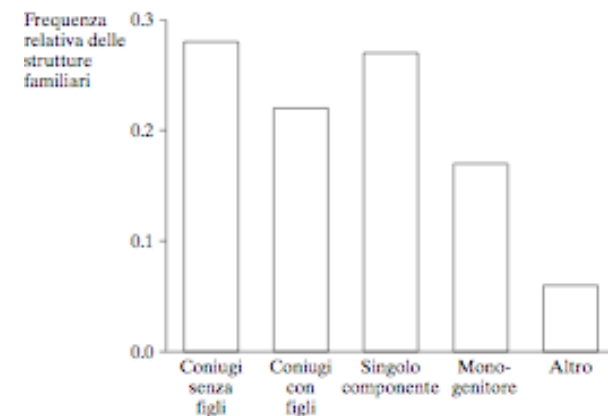
Descrizione dei dati (Statistica descrittiva)

L'obiettivo della statistica descrittiva è quello di sintetizzare i dati per rendere fruibili le informazioni in essi contenute

Tabelle e grafici sono utili per sintetizzare tutti i tipi di dati

Famiglia	Numerosità	Proporzione	%
Coppie sposate con figli	24.1	0.22	22
Coppie sposate senza figli	31.1	0.28	28
Monogenitore	19.1	0.17	17
Singolo componente	30.1	0.27	27
Altre tipologie	6.7	0.06	6
Totale	111.1	1.00	100

Fonte: U.S. Census Bureau, 2005 Am.Comm.Survey, Tav. B11001, C11003.



Come si ottengono tabelle e grafici?

Tipi di caratteri (variabili statistiche)

QUESTIONARIO

D1 Guardi serie TV?

1. Sì
2. No

D2 Ti piace andare al cinema?

1. Molto
2. Abbastanza
3. Né tanto né poco
4. Poco
5. Per niente

D3 Età (anni)

D4 Quante volte vai al cinema in un mese?

QUALITATIVI

QUANTITATIVI

I caratteri **qualitativi** presentano diverse modalità o categorie e non sono *misurabili*;

I caratteri **quantitativi** sono espressi da conteggi o misurati su scala numerica (le modalità sono espresse in forma numerica)

I dati qualitativi possono esprimere caratteri *nominali* e caratteri *ordinali* a seconda che sia o meno possibile stabilire un ordinamento tra le modalità.

Esercizio

Individuare il tipo di carattere:

- Strumento musicale suonato
- Genere di film
- Codice postale
- Numero di chiamate ricevute da un operatore di call center
- Soddisfazione dei clienti di Ryanair
- Tempo medio giornaliero impiegato per raggiungere il luogo di lavoro a Trieste

Distribuzione di un carattere

Dopo aver acquisito e controllato i dati si passa alla loro sintesi e descrizione

Distribuzione unitaria: elenco, unità per unità, delle modalità di una variabile osservate nel campione

La variabile assume **una** modalità in corrispondenza di **ogni** unità statistica

- La variabile età assume valore 21 per l'unità 5
- La distribuzione unitaria del carattere sesso è:
 - F, F, M, F, F, F, M, F

Matrice dei dati: distribuzioni unitarie delle variabili (dati di base)

colonne: elenco delle modalità di ciascuna variabile osservata nel campione

righe: unità osservate

Unità	Sesso	Età	Statura	Colore occhi
1	F	24	163	Marrone
2	F	21	165	Azzurri
3	M	34	185	Azzurri
4	F	22	164	Marroni
5	F	21	167	Marroni
6	F	22	175	Verdi
7	M	24	178	Verdi
8	F	21	155	Marroni

Distribuzione di frequenze

La **frequenza assoluta** di una modalità è il numero di volte che questa viene osservata nel campione

- La frequenza assoluta della modalità “Monogenitore” per la variabile Famiglia è 19,1 milioni

La **distribuzione di frequenze** associa alla distribuzione di una variabile le frequenze osservate

- Si dice **semplice (univariata)** se riferita ad un sola variabile, **doppia** se riferita a due variabili, **multipla** a più di una variabile.

Famiglia	Numerosità	Proporzione	%
Coppie sposate con figli	24.1	0.22	22
Coppie sposate senza figli	31.1	0.28	28
Monogenitore	19.1	0.17	17
Singolo componente	30.1	0.27	27
Altre tipologie	6.7	0.06	6
Totale	111.1	1.00	100

Fonte: U.S. Census Bureau, 2005 Am.Comm.Survey, Tav. B11001, C11003.

Costruiamo le distribuzioni di frequenze per le variabili del nostro campione

Unità	Sesso	Età	Statura	Colore occhi
1	F	24	163	Marrone
2	F	21	165	Azzurri
3	M	34	185	Azzurri
4	F	22	164	Marroni
5	F	21	167	Marroni
6	F	22	175	Verdi
7	M	24	178	Verdi
8	F	21	155	Marroni

Sesso	Numerosità
F	6
M	2
	8

Età	Numerosità
21	3
22	2
24	2
34	1
	8

- Qual è la modalità osservata più numerosa della variabile Età?
[21]
- Quale variabile è più difficile sintetizzare?
[le variabili quantitative o le qualitative con molte modalità]
- Quale delle due rappresentazioni dei dati raccolti offre maggiori informazioni?
[la distribuzione unitaria]
- Da quale rappresentazione è più semplice leggere informazioni?
[le distribuzioni di frequenze]

Suddivisione in classi

Quando la variabile presenta molte modalità distinte è utile procedere ad una divisione in classi

Non esiste una regola unica per la suddivisione:

- è una scelta soggettiva, dipende dal contesto e per questo deve essere motivata
- si perdono informazioni al prezzo di una maggiore leggibilità dei dati osservati

Se la variabile è qualitativa si possono accorpare le modalità seguendo uno specifico criterio (ad esempio un livello superiore di gerarchia: comuni -> province -> regioni)

Se la variabile è quantitativa la **suddivisione in classi** ci porta ad un livello ordinale

- Le classi possono avere **ampiezza costante** o **diversa**
- Se il **numero delle classi** è troppo piccolo, rischiamo di sintetizzare troppo e perdere troppa informazione viceversa, se il numero delle classi è troppo alto manteniamo più informazione ma rischia di essere poco leggibile (troppi dettagli)
- Le classi devono essere **disgiunte (mutualmente esclusive)** e devono includere tutte le possibili modalità della variabile

Suddivisione in classi

L'ampiezza delle classi può essere calcolata come:

- $ampiezza = \frac{\text{valore massimo} - \text{valore minimo}}{\text{numero delle classi}}$
 - il minimo e massimo valore osservato non devono coincidere con l'estremo inferiore della prima classe e con l'estremo superiore dell'ultima
 - l'ampiezza ottenuta va approssimata ad un numero intero (es. 9,7 -> 10)

Esempio:

I dati osservati variano tra 11,2 e 98,6 e si vogliono suddividere in 9 classi:

$$ampiezza = \frac{98,6 - 11,2}{9} \approx 10$$

Scegliamo come valore iniziale per la prima classe 10 (così da non farlo coincidere con 11,2), avremo [10, 20), [20, 30), ..., [90,100) oppure 10 -19, 20 - 29, ..., 90 - 99

La leggibilità dei dati è la priorità!

Frequenze relative e percentuali

La **frequenza relativa** è la frequenza assoluta divisa per il numero totale di unità osservate

- È un numero compreso tra 0 e 1
- La somma delle frequenze relative di una variabile è uguale a 1

La **frequenza percentuale** è la frequenza relativa moltiplicata per 100

Età	Numerosità	Freq Relativa	Percentuale
21	3	$3/8 = 0,375$	$3/8 * 100 = 37,5\%$
22	2	$2/8 = 0,25$	$2/8 * 100 = 25\%$
24	2	$2/8 = 0,25$	$2/8 * 100 = 25\%$
34	1	$1/8 = 0,125$	$1/8 * 100 = 12,5\%$
	8	1	100

Le frequenze relative o percentuali sono utili per **confrontare frequenze** da campioni di diversa numerosità poichè non dipendono dalla numerosità del campione

Esempio: dalle distribuzioni di frequenze assolute dei due campioni qui sotto sembra che la modalità 1 sia **più presente** nel secondo gruppo: (gruppo1: $x_1 = 2$; gruppo2: $x_2 = 12$)

Gruppo 1	Numerosità
x_1	2
x_2	4
x_3	8
	14

Gruppo 2	Numerosità
x_1	12
x_2	46
x_3	32
	90

Considerando però le frequenze percentuali otteniamo che in realtà la modalità 1 è più presente nel gruppo 1

$$\text{gruppo1: } p_1 = \frac{2}{14} * 100 = 14,29\%$$

$$\text{gruppo2: } p_1 = \frac{12}{90} * 100 = 13,33\%$$